

# Bibliotecas, repositorios y otras bases de datos Web

## Una nueva realidad ...

---

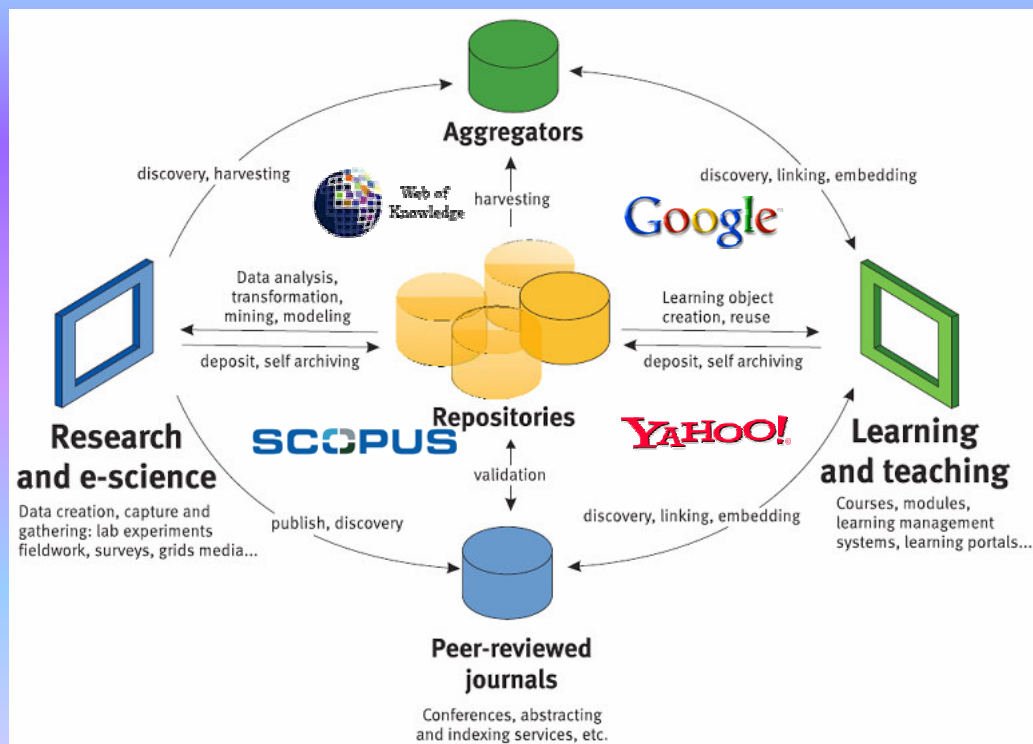


- Bibliotecas digitales “virtuales”

*Colecciones de documentos distribuidas*

- **Web visible:** Auto-archivo, depósitos de documentos, buscadores de área
- **Web invisible:** Bases de datos, repositorios institucionales y temáticos, revistas electrónicas
- **Internet privada:** Consorcios nacionales, regionales e institucionales

## ... para un nuevo modelo



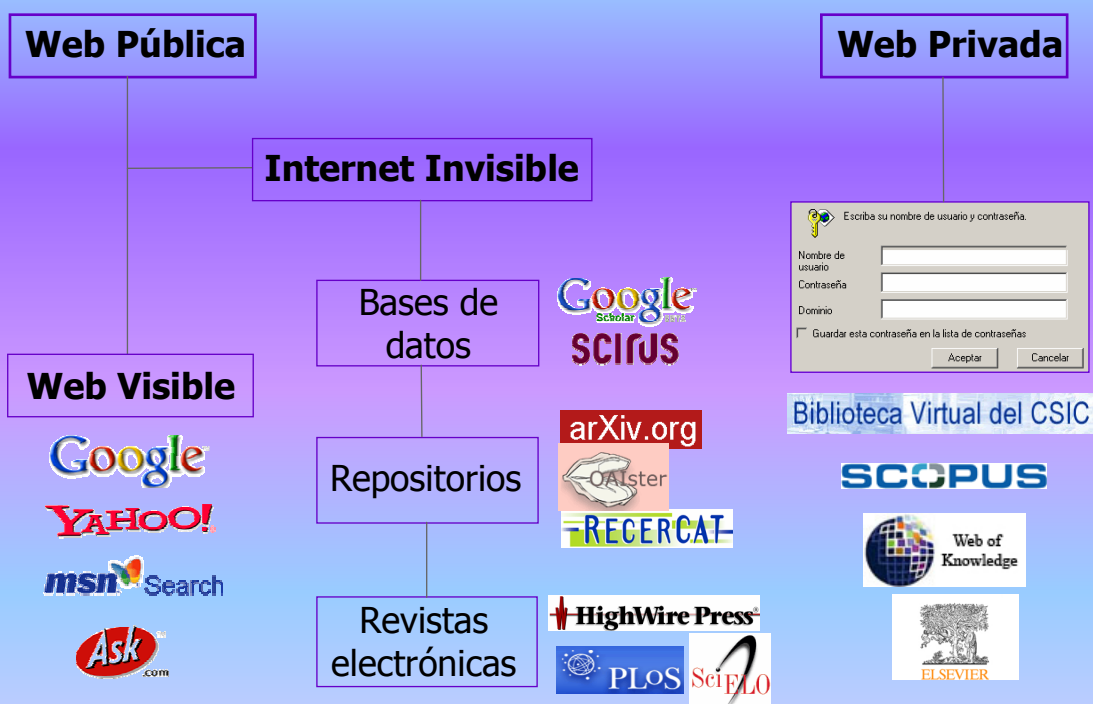
## Tres escenarios

---



- **Web pública**
  - **Web visible**  
Información recogida en los motores de búsqueda
  - **Web invisible**  
Bases de datos bajo pasarelas no indizadas por los motores de búsqueda
- **Internet privada**  
Se requiere subscripción o identificación vía IP

# Recursos académicos en el Ciberespacio



# Análisis cuantitativo (I)

---



- **Web visible**

Objetos: Ficheros ricos

(autoarchivo, páginas personales e institucionales, depósitos)

- Adobe Acrobat (.pdf): Artículos, informes, monografías
- MS Word (.doc): material definitivo y borradores
- Postscript (.ps): Artículos (Física, Matemática, Ingeniería, ...)
- MS Powerpoint (.ppt): congresos, docencia

- **Método de acceso: Motores de búsqueda**

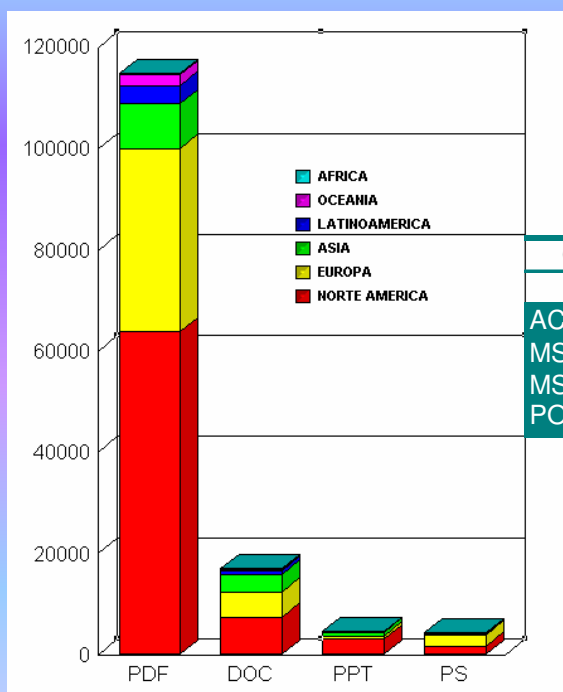
Rápido y flexible

Alta visibilidad e impacto (acceso gratuito?)

Sin control documental

Cobertura escasa e irregular

# Ficheros ricos



CONTRIBUCIÓN ACADEMICA (Google, miles)				
FORMATOS	GLOBAL	UNIVERSIDADES		
ACROBAT <b>PDF</b>	608.000	114.849	18,9%	
MS WORD <b>DOC</b>	119.000	16.898	14,2%	
MS POWERPOINT <b>PPT</b>	22.900	4.447	19,4%	
POSTSCRIPT <b>PS</b>	13.600	4.193	30,8%	

(Datos propios, Julio 2006)

# Otros modelos en la Web visible



Webpags	España I+D+I
8.156	Agricultura, Pesca y Alimentación
17.995	Ciencias Biológicas
54.798	Ciencias de la Salud
16.275	Ciencias de la Tierra
35.762	Ciencias Físicas y Exactas
13.070	Ciencias Químicas
23.897	Ciencias Sociales y Humanidades
28.781	Energía
49.084	Industria y Tecnología
254	Medidas y Normas
19.976	Medio Ambiente
8.854	Políticas Tecnológicas de I+D
556.455	TOTAL

Noviembre 2006

**Buscador España I+D+I**

Buscador especializado en Ciencia y Tecnología.

España I+D+I presenta actualmente **307 centros indexados**.

- 75 Organismos miembros del Sistema madri+d
- 44 Centros de investigación
- 83 Centros de cultura científica y participación
- 32 Centros de innovación
- 19 Empresas innovadoras
- 32 Organismos públicos de planificación, financiación y evaluación de la política científica
- 22 Portales especializados en Ciencia y Tecnología

Tras la búsqueda, los resultados obtenidos se clasifican automáticamente en taxonomías utilizando sistemas avanzados de ingeniería lingüística.

Puede realizar sus consultas navegando a través de la taxonomía, realizando búsquedas a texto libre o bien combinando ambas posibilidades.

■ **Solicitud de indexación.**

■ **Visite la ayuda para optimizar sus búsquedas.**

<http://buscador.madrimasd.org/BuscadorMadrimasd/default.asp>



# ¡No es tan difícil!



**Google Co-op BETA Custom Search Engine**

Harness the power of Google search to create a free Custom Search Engine that reflects your own websites that you want searched - and integrate the search box and results into your own website.

**Build and customize your own search engine**

- Specify the sites you want to include in searches.
- Place a search box and search results on your website.
- Customize the look and feel to match your website.
- Invite your community to contribute to the search engine.
- Make money from relevant ads in your search results.
- Learn more: [FAQ](#) and [featured examples](#).

**Already have a Custom Search Engine?**

Check out your search engine's homepage and control panel on your [My search engines](#) page.

**Get started**

1. Specify your search engine
2. Try it out

[Create a Custom Search Engine](#)

**lbr**

**Custom Search Engines via Google Co-op**

CH E & librarianship / technology / resources / [businesslibrarian\\_email](#)

posted 2006.11.09 Thursday

As soon as I saw what it was about and what it could do, I knew what I wanted to try to do with it. [I've previously discussed](#) the [Directory of Open Access Journals \(DOAJ\)](#) on this blog. DOAJ is truly an incredibly valuable collection of resources -- and it does what it set out to do very well, which is to be a [directory of open-access journals](#) -- not a full-text index. But it's hard to browse that fantastic directory without wishing for a magic search box that could perform a search across that entire collection. DOAJ has actually started down that road on their own -- of the 2450 journals in the directory today, 721 can be searched at the article level using the [Find Articles tool](#) on the DOAJ site.

But with the advent of CSE, it seemed to me that we might be able to create that "magic search box" another way. So I went to the DOAJ site and grabbed the journal metadata in CSV format [as described in the FAQ](#) (note that it's licensed [CC-BY-SA-1.0](#)). Then, with a few Excel hacks, I was able to parse out a list of domains from that list that hosted English-language DOAJ journals. I carved out the domain names, added a slash and asterisk at the end to indicate I wanted everything in that domain, and dropped them in the batch upload box for CSE. And just like that -- we have what you might call an [early prototype of a "magic search box"](#) for English-language DOAJ journals.

DOAJ English 1:

**Catorze .blog**

[Qué es](#) [Contenidos](#) [Servicios web](#) [Blog](#) [Contacto](#)

**Buscador de Blogs de ByD (Google Co-op)**

La presente página es una prueba de búsqueda restringida a blogs sobre Biblioteconomía y documentación usando Google Co-op. ([leer explicación](#))

**BUSCOPIUM**

Buscador medieval

[Sitios nuevos](#) [Sitios favoritos](#) [XML - RSS](#) [Sugerir enlace](#)

**BUSCOPIUM**

**ARMAS Y EJÉRCITOS**  
Campañas militares - Batallas, Órdenes Militares, Fortalezas, Armamento, ...

**FUENTES DOCUMENTALES**  
Crónicas, Selección de textos, Cronologías, Bibliografías, Cartularios, Actas, Epistolarios, ...

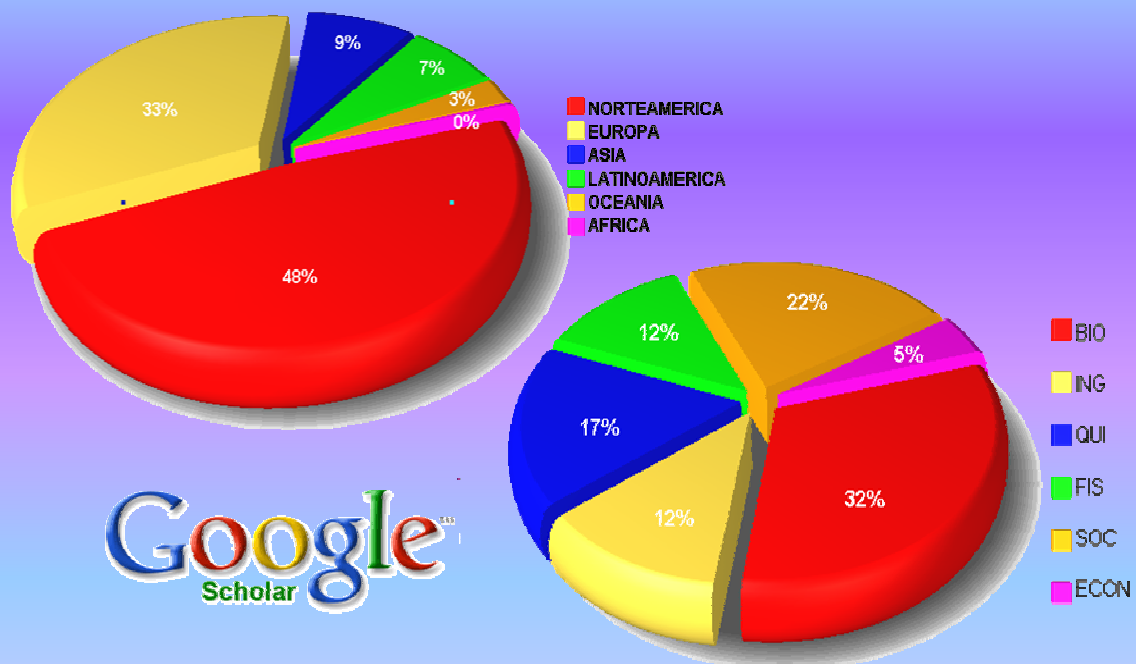
## Análisis cuantitativo (II)

---



- Web invisible y **privada**
  - Referencias bibliográficas:
    - Intermediarios
      - Google Scholar
      - MSN Live Academic
      - Elsevier Scirus
      - ISI Web Citation Index**
      - Scopus**

# Google Académico



Objetos en dominios universitarios (Datos propios, Julio 2006)

# Live Academic



LIVE	SCHOLAR	DOMINIOS	
69,72%	39,00%	com+org+net+info...	Editoriales, Asociaciones
14,06%	6,37%	edu	Universidades
8,89%	0,89%	uk	Editoriales/Universidades
1,40%	2,29%	de	
0,63%	7,17%	fr	Universidades
0,51%	0,42%	nl	Editoriales
0,46%	18,92%	us+gov+mil	Informes
0,44%	0,77%	ca	
0,41%	0,33%	ch	
0,36%	0,63%	au	
0,29%	0,47%	it	
0,28%	1,12%	jp	
0,28%	0,33%	se	
0,23%	0,52%	es	
0,19%	0,27%	be	
0,18%	0,27%	at	
0,16%	0,28%	dk	
0,16%	0,29%	fi	
0,14%	0,18%	il	
0,10%	0,16%	gr	
0,09%	0,30%	kr	
0,09%	0,14%	ie	
0,09%	1,93%	br	
0,09%	0,26%	pt	
0,08%	0,28%	cz	
2.739.120	8.245.872	total de ficheros	

Web

Imágenes

Noticias

Académico

Editoriales

Revistas

Actas de Congresos

118

4.371

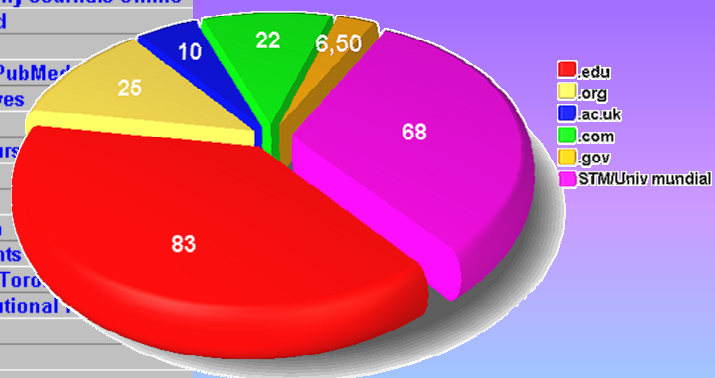
2.086

Octubre 2006

# Scirus



13.000.000	artículos	BioMed Central
6.600.000	artículos	ScienceDirect
368.000	artículos	PubMed Central
330.000	artículos	Scitation
212.000	artículos	Institute of Physics Publishing
65.500	artículos	Crystallography Journals Online
35.500	artículos	Project Euclid
8.800	artículos	SIAM
17.300.000	referencias	Medline via PubMed
390.000	documentos	Digital Archives
180.000	documentos	RePEc
63.000	documentos	MIT OpenCourseWare
11.000	documentos	DiVa
6.500	documentos	WaY
6.000	documentos	Caltech Coda
5.200	documentos	Organic Eprints
4.400	documentos	University of Toronto
2.200	documentos	HKUST Institutional Repository
1.700	documentos	CURATOR
600	documentos	PsyDok
363.500	e-prints	ArXiv.org
2.600	e-prints	Cogprints
13.000.000	patentes	LexisNexis
12.000	informes técnicos	NASA
237.500	tesis	NDLTD



Millones de páginas Web

Julio 2006

# Web Citation Index



378.523	e-prints	90,45%	ARXIV ORG E PRINT ARCHIVE
7.999	e-prints	1,91%	DSPACE MIT
4.305	e-prints/informes	1,03%	CALTECH
2.995	e-prints	0,72%	AUSTRALIAN NATIONAL UNIVERSITY
2.805	e-prints	0,67%	BIOLINE INTERNATIONAL
2.686	e-prints	0,64%	E LIS
2.640	informes	0,63%	DARTMOUTH COLLEGE COMPUTER SCIENCE
2.544	e-prints	0,61%	EPRINTSUQ
1.843	tesis	0,44%	UNIVERSITY OF HELSINKI
1.675	e-prints	0,40%	COGPRINTS COGNITIVE SCIENCES
1.332	informes	0,32%	BUREAU OF LABOR STATISTICS
1.287	informes	0,31%	OECD
1.207	informes	0,29%	COMPUTER LABORATORY UNIVERSITY OF CAMBRIDGE
942	e-prints	0,23%	SCHOLARLYCOMMONS PENN
869	e-prints	0,21%	UNIVERSITY OF TORONTO
641	e-prints	0,15%	QUEPRINTS CRANFIELD UNIVERSITY
598	informes	0,14%	COMPUTER SCIENCE IOWA STATE UNIVERSITY
556	e-prints	0,13%	EUROPEAN RESEARCH
539	e-prints	0,13%	EDINBURGH RESEARCH
439	e-prints	0,10%	DSPACE DREXEL UNIVERSITY
407	e-prints	0,10%	ESPACE CURTIN
345	e-prints	0,08%	DSPACE UNIVERSITY OF ROCHESTER
322	e-prints	0,08%	GLASGOW
209	e-prints	0,05%	DSPACE UNIVERSITY OF
202	e-prints	0,05%	UNIVERSITY OF NOTTINGHAM
194	informes	0,05%	EUROPEAN CENTRE FOR
177	e-prints	0,04%	MATHEMATICAL INSTITUTE
143	e-prints	0,03%	OXFORD EPRINTS
54	e-prints	0,01%	DIGITALCOMMONS UTE
27	e-prints	0,01%	HISTORY THEORY OF PSYCHOLOGY
5	e-prints	0,00%	INDIAN INSTITUTE OF SCIENCE RESEARCH PUBLICATIONS

Servicio gratuito proporcionado por la FECYT y el MEC

ISI Web of Knowledge<sup>SM</sup>



Web Citation Index

Octubre 2006

# Scopus



15.000	Títulos con revisión por pares
500	Revistas Open Access
700	Actas de Congresos
600	Informes comerciales
125	Series de monografías
28.000.000	Resúmenes
245.000.000	Referencias bibliográficas

SCOPUS™

## WebCites

Right hand panel of Abstract & Refs page

### Cited By - Web Sources

[26 times](#)

Covered web sources: University repositories (e.g. MIT, DiVA, Caltech), theses & dissertations.

MIT - Open Courseware from the Massachusetts Institute of Technology (MIT)  
NDLTD - Networked Digital Library of Theses and Dissertations (170,000 documents remaining 30% will be done by the end of 2006)  
DiVA - a collection of Institutional repositories from Scandinavian universities  
University of Toronto - The institutional repository of University of Toronto  
CalTech - The institutional repository from the California Institute of Technology.

Noviembre 2006

15

## Otros Repositorios

---



- **Institucionales**

- Artículos (E-prints Complutense)

- Tesis, tesinas (TDCAT)

- **Temáticos**

- Pre-prints (arXiv, E-LIS)

- Artículos (DOIS, RePEC)

- Artículos + Citas (Scielo)

- Citas + Artículos (CiteSeer)

- BBDD bibliográficas con enlaces (Dialnet)

- **Revistas electrónicas**

- Directorios + Bases de datos (Highwire)



# Enlaces "activos" (I)



**Query-Free News Search (2003)** ([Make Corrections](#)) ([2 citations](#))  
Monika Henzinger, Bay-Wei Chang, Brian Milch, Sergey Brin

**CiteSeer.IST** @ University of Zurich  
Scientific Literature Digital Library @ Department of Informatics

[Home/Search](#) [Bookmark](#) [Context](#) [Related](#)

View or download:  
[henzinger.com/monik\\_p707henzinger.ps](http://henzinger.com/monik_p707henzinger.ps)  
[berkeley.edu/~milch/papers\\_www2003.ps](http://berkeley.edu/~milch/papers_www2003.ps)  
Cached: [PS.gz](#) [PS](#) [PDF](#) [Image](#) [Update](#) [Help](#)

From: [henzinger.com/mon\\_publications](http://henzinger.com/mon_publications) ([more](#))  
([Enter author homepages](#))

**Abstract:** Many daily activities present information in the form of a stream of text, and often people can benefit from additional information on the topic discussed. TV broadcast news can be treated as one such stream of text; in this paper we discuss finding news articles on the web that are relevant to news currently being broadcast. ([Update](#))

**Cited by:** [More](#)  
The Anatomy of a News Search Engine - Gulli Dipartimento Di (2005) ([Correct](#))  
Universit  di Pisa - Gulli, Ferragina ([Correct](#))

**Active bibliography (related documents):** [More](#) [All](#)  
[0.5](#) Context-Based Similarity Applied to Retrieval of Relevant Cases - Jurisica (1994) ([Correct](#))  
[0.0](#) Collaborative Recommender Agents Based on Case-Based Reasoning. - Montaner (2003) ([Correct](#))  
[0.0](#) A System For Automatic Personalized Tracking of.. - Bollacker, Lawrence. (1999) ([Correct](#))

**Similar documents based on text:** [More](#) [All](#)  
[1.2](#) Challenges in Web Search Engines - Henzinger, Motwani, Silverstein (2002) ([Correct](#))  
[0.6](#) Least Common Ancestors in Trees and Directed Acyclic.. - Bender, Farach-Colton, (2001) ([Correct](#))  
[0.6](#) When Experts Agree: Using Non-Affiliated Experts to Rank.. - Bharat, Mihaila (2001) ([Correct](#))

**BibTeX entry:** ([Update](#))

Monika Henzinger, Bay-Wei Chang, Brian Milch, and Sergey Brin. Queryfree news search. In WWW12, pages 1--10, 2003. <http://citeseer.ifi.unizh.ch/henzinger03queryfree.html> [More](#)


```
@misc{henzinger03queryfree,
  author = "M. Henzinger and B. Chang and B. Milch and S. Brin",
  title = "Queryfree news search",
  text = "Monika Henzinger, Bay-Wei Chang, Brian Milch, and Sergey Brin. Queryfree news search. In WWW12, pages 1--10, 2003.",
  year = "2003",
  url = "citeseer.ifi.unizh.ch/henzinger03queryfree.html" }
```

**Citations (may not include all citations):**  
166 Transformation-based error-driven learning and natural langu.. - Brill - 1995  
23 What can you do with a web in your pocket - Brin, Motwani et al. - 1998  
18 Temporal summaries of news topics - Allan, Gupta et al. - 2001  
16 Information access in context (context) - Budzik, Hammond et al. - 2001  
15 Domain-specific keyphrase extraction - Frank, Paynter et al. - 1999  
11 Learning user information interests through the extraction o.. (context) - Krulwich, Burkey  
2 Query-free information retrieval - Hart, Graham - 1997  
1 EIA-746-A: Transport of internet uniform resource locator (context) - Alliance - 1998  
1 Interact dying of neglect (context) - Davis - 1997

**Documents on the same site (<http://www.henzinger.com/monika/publications.html>):** [More](#)  
[Analysis of a Very Large AltaVista Query Log - Silverstein \(1998\)](#) ([Correct](#))

## Enlaces "activos" (II)



**DIALNET**  
Bienvenido, Usuario no registrado

[Nueva Búsqueda](#) [Tesis](#) [Conectar](#) [Alta Usuario](#) [Ayuda](#)

**Tomás Baiget** [sugerencia/errata](#) [Volver](#)

Revistas ( 19 artículos ) Monografías Colectivas ( 2 artículos )

**Web** <http://www.sarenet.es/baiget/>

**Artículos de Revistas**  
**Programa work-flow para un servicio de consultas a medida** / Tomás Baiget  
**En:** *El profesional de la información*, ISSN 1386-6710, Vol. 15, Nº 5, 2006 (Ejemplar dedicado a: Intranets), pags. 364-372  
[Resumen] [ [Texto Completo Artículo](#) ]  
**Les quinze coses que més han influït en la biblioteconomia i la documentació en els darrers quinze anys** / Tomás Baiget  
**En:** *Item: Revista de biblioteconomia i documentació*, ISSN 0214-0349, Nº. 43, 2006, pags. 69-90  
[Resumen]  
**La publicación médica en España : crónica del curso** / Tomás Baiget  
**En:** *El profesional de la información*, ISSN 1386-6710, Vol. 14, Nº 5, 2005, pags. 391-395  
**Tendències i efemèrides 2004** / Tomás Baiget  
**En:** *Bibliodoc: Anuari de Biblioteconomia, Documentació i Informació = Anuario de Biblioteconomía, Documentación e Información = Library and Information Sciences Yearbook*, Nº. 2004, 2005, pags. 149-174  
**Uso de recursos de información electrónicos 2003-2004** / Concha Alvaro, Dirk Lens, Juan Carlos Martín, Carlos Tejada, Tomás Baiget, Alice Keefer Riva, Elea Giménez Toledo  
**En:** *El profesional de la información*, ISSN 1386-6710, Vol. 13, Nº 5, 2004, pags. 386-392  
[Resumen]  
**Icsep 2004 : II Taller latinoamericano: recursos y posibilidades de la publicación electrónica** / Tomás Baiget  
**En:** *El profesional de la información*, ISSN 1386-6710, Vol. 13, Nº 3, 2004, pags. 232-241

## Justificación

---



- **Visibilidad**
  - Geográfica (países en vías de desarrollo)
  - Socioeconómica (audiencias más amplias)
- **Acceso**
  - Universal, iniciativas "Open Access"
- **Tecnológica**
  - Interconexión "profunda"
- **Política**
  - Herramienta de evaluación
  - Identificando la "brecha digital"

# R anking Europeo de Universidades en la Web

Julio '06















inicio pa ses del mundo ranking mundial ranking europeo ranking latino americano ranking espa ol

<b>Datos</b>
Top 500 Univ. Europeas
Top 100 por Pa�s
Top 100 I+D Institutos
Lista de Centros I+D
Especiales
Buenas Pr�cticas

<b>An�lisis Comparativo</b>
Productividad
Visibilidad
Impacto
Metodolog�a

<b>Cat�logo</b>
Universidades por Pa�s
Centros I+D por Pa�s

<b>Informaci�n</b>
Metodolog�a
Glosario
Blog
Enlaces
Contacta con nosotros
Mapa del Sitio

Top 500 Universidades Europeas								
1-100   101-200   201-300   301-400   401-500   Universidades 1 a 100 de 500								
RANKING		UNIVERSIDAD	PA�S	TAMA�O	VISIB.	POSICI�N		SCHOLAR
EUROPA	MUNDO					FICHEROS	RICOS	
1	19	UNIVERSITY OF CAMBRIDGE		25	24	38		48
2	22	UNIVERSITY OF OXFORD		15	29	44		64
3	37	SWISS FEDERAL INSTITUTE OF TECHNOLOGY ZURICH		22	48	77		33
4	44	UNIVERSITY OF EDINBURGH		77	44	34		28
5	51	UNIVERSITY OF OSLO		51	69	18		122
6	59	LINKOPING UNIVERSITY		26	84	116		74
7	63	UNIVERSITY OF HELSINKI		69	91	15		118
8	65	ROYAL INSTITUTE OF TECHNOLOGY		93	97	39		72
9	67	UNIVERSITY COLLEGE LONDON		105	73	112		81
10	69	FREE UNIVERSITY OF BERLIN		68	88	107		111
11	75	NORWEGIAN UNIVERSITY OF SCIENCE AND TECHNOLOGY		62	123	22		144
12	79	UPPSALA UNIVERSITY		56	147	30		66
13	82	UNIVERSITY OF HAMBURG		52	127	110		91
14	83	UTRECHT UNIVERSITY		106	129	47		40
15	87	UNIVERSITY OF VIENNA		114	93	82		147

# Universidades españolas y CSIC



RANKING	EUROPE	WORLD	UNIVERSITY	COUNTRY	SIZE	POSITION		
						COUNTRY	SIZE	SCHOLAR
1	19		UNIVERSITY OF CAMBRIDGE		25	24	38	48
2	22		UNIVERSITY OF OXFORD		15	29	44	64
3	37		SWISS FEDERAL INSTITUTE OF TECHNOLOGY ZURICH		22	48	77	33
42	147		UNIVERSITY COMPLUTENSE MADRID		82	231	253	17
84	248		AUTONOMOUS UNIVERSITY OF BARCELONA		303	315	288	71
86	255		POLYTECHNIC UNIVERSITY OF CATALONIA		255	344	273	139
90	259		POLYTECHNIC UNIVERSITY OF MADRID		278	297	285	297
91	262		UNIVERSITY OF BARCELONA (UB.ES)		336	348	167	120

## Ranking Universidades Europeas

## Ranking Centros I+D

INSTITUTE	COUNTRY	POSITION			RICH FILES	SCHOLAR
		COUNTRY	SIZE	VISIBILITY		
NATIONAL INSTITUTES OF HEALTH		3	2	12	1	
2 NATIONAL AERONAUTICS AND SPACE ADMINISTRATION		4	4	2	4	
8 CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE CNRS		13	15	13	5	
8 MAX PLANCK GESELLSCHAFT		9	14	14	16	
13 CONSIGLIO NAZIONALE DELLE RICERCHE		17	24	10	32	
33 CONSEJO SUPERIOR DE INVESTIGACIONES CIENTIFICAS		54	61	24	27	

## ... ¿el futuro?



- **Más contenidos**
  - Páginas personales, blogs, wikis, y otros medios informales
  - Integración de las revistas en los repositorios
- **Más estructura semántica**
  - Valor añadido mediante interconexión hipertextual (super-citas)
  - Web semántica y Web 2.0: Metadatos, etiquetas XML, ontologías RDF
- **Menos volatilidad**
  - Preservación distribuida

## Conclusiones

---



- Bibliotecas digitales: Escenario rico y diverso
  - Publicación formal: Impacto positivo de las iniciativas *Open Access* (auto-archivo, repositorios institucionales, repositorios temáticos)
  - Publicación informal: Fases de la actividad académica reveladas en la Web
  - Nuevos intermediarios: Acceso universal, nuevos retos para el documentalista
  - Formatos con mayor valor añadido: Web Semántica, interconexión, anotación distribuida, ... Web 2.0, 3.0

¡Gracias!

---



¿Preguntas?

Más información:

Sede Web	<a href="http://internetlab.cindoc.csic.es">http://internetlab.cindoc.csic.es</a>
Ranking de Universidades	<a href="http://www.webometrics.info">http://www.webometrics.info</a>
Revista-e Cybermetrics	<a href="http://www.cindoc.csic.es/cybermetrics">http://www.cindoc.csic.es/cybermetrics</a>